British Journal of Psychology (2019), 110, 576–593

The British Psychological Society

www.wileyonlinelibrary.com

How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices

Nadine Lavan[1,2]*, Luke F. K. Burston[1] and Lúcia Garrido[2]

[1]Department of Psychology, Royal Holloway, University of London, Egham, UK
[2]Division of Psychology, Department of Life Sciences, Brunel University, London, UK

Our voices sound different depending on the context (laughing vs. talking to a child vs. giving a speech), making within-person variability an inherent feature of human voices. When perceiving speaker identities, listeners therefore need to not only 'tell people apart' (perceiving exemplars from two different speakers as separate identities) but also 'tell people together' (perceiving different exemplars from the same speaker as a single identity). In the current study, we investigated how such natural within-person variability affects voice identity perception. Using voices from a popular TV show, listeners, who were either familiar or unfamiliar with this show, sorted naturally varying voice clips from two speakers into clusters to represent perceived identities. Across three independent participant samples, unfamiliar listeners perceived more identities than familiar listeners and frequently mistook exemplars from the same speaker to be different identities. These findings point towards a selective failure in 'telling people together'. Our study highlights within-person variability as a key feature of voices that has striking effects on (unfamiliar) voice identity perception. Our findings not only open up a new line of enquiry in the field of voice perception but also call for a re-evaluation of theoretical models to account for natural variability during identity perception.

Voices are highly variable. The same person can sound very different depending on the speaking context: For example, we modulate pitch, speech rate, and speaking style depending on whether we are giving a public lecture, talking to a friend, or singing (Kreiman, Park, Keating, & Alwan, 2015; Lavan, Burton, Scott, & McGettigan, 2018). Thus, within-person variability is an inherent feature of the human voice that we encounter in all of our interactions. Despite its ubiquity, within-person variability poses challenges to identity perception from vocal signals: Listeners do not only have to tell different voices apart, but they also need to generalize percepts of identity across substantial within-person variability to maintain a level of constancy in identity perception (i.e., 'telling people together'; see Burton, 2013 for faces). Arguably, being able to 'tell people together' can only be reliably achieved for familiar voices (Jenkins, White, Van Montfort, & Burton, 2011 for faces) – we may need to have learned how a specific voice varies to not mistake the substantial inherent within-person variability as between-person variability.

Traditionally, studies of how we recognize people from their voices have explicitly controlled for and thus minimized within-person variability: The experimental stimuli

**Correspondence should be addressed to Nadine Lavan, Department of Psychology, Royal Holloway University of London, Egham, Surrey TW20 0EX, UK (email: nadine.lavan.2013@rhul.ac.uk).*

DOI:10.1111/bjop.12348

used tend to be carefully selected recordings of vowels, words, or short sentences produced in neutral intonation, mostly recorded in a single recording session. This approach has allowed us to gain insights into how we tell people apart via the distinguishing features of individual voices. It has, however, at the same time restricted our understanding of voice identity perception to this particular set of contexts, neglecting the study of the perceptual mechanisms that we use to compute stable and consistent representations of familiar voices despite the substantial within-person variability (Lavan, Scott, & McGettigan, 2016; Lavan *et al.*, 2018). Similar issues have recently been highlighted for the face identity processing literature (Burton, 2013; Burton, Kramer, Ritchie, & Jenkins, 2016), opening up a fruitful new line of enquiry in this field.

Below, we first review the few studies that have started to investigate how within-person variability and familiarity affect voice identity processing. We then summarize findings from the face perception literature showing striking interactions between familiarity and within-person variability, and propose that adapting a paradigm from this field (Jenkins *et al.*, 2011) to voices will allow us to shed new light on mechanisms of voice identity processing.

Effects of within-person variability in voice identity processing

In the presence of within-person variability, listeners make more errors during identity perception. For example, listeners are less accurate at correctly matching speakers across pairs of sentences produced in different languages compared to when pairs include the same language (Wester, 2012; Zarate, Tian, Woods, & Poeppel, 2015). Furthermore, linguistic (dis)similarity of stimuli affects speaker discrimination performance in a top-down fashion: Identities can be more accurately discriminated from pairs of stimuli that are semantically or phonetically related, such as 'day-dream' or 'day-bay', than from linguistically unrelated stimuli, such as 'day-bee' (Narayan, Mak, & Bialystok, 2017). Similarly, listeners fail to reliably discriminate between unfamiliar identities when making judgements for pairs of disguised and undisguised voices (e.g., hypernasal voice vs. neutral voice; Reich & Duke, 1979), across different vocalizations (e.g., vowels vs. laughter; Lavan *et al.*, 2016), and across sung versus spoken words (Peynircioğlu, Rabinovitz, & Repice, 2017). In forensic contexts, studies of earwitness' judgements report that listeners' ability to identify a voice from a line up decreases when vocal variability (e.g., through changes in emotional tone) is introduced between study and test (Read & Craik, 1995; Saslove & Yarmey, 1980). Even when listeners are familiar with a voice, they are unable to accurately recognize known individuals speaking in falsetto voice versus modal ('normal') voice (Wagner & Köster, 1999).

Despite this growing body of literature, current models of voice processing do not explicitly account for within-person variability. For example, prototype models are often used as a theoretical basis to map out how different identities are encoded and how they may relate to each other (Latinus & Belin, 2011; Latinus, McAleer, Bestelmeyer, & Belin, 2013; Lavner, Rosenhouse, & Gath, 2001; Papcun, Kreiman, & Davis, 1989; see also Maguinness, Roswandowitz, & Von Kriegstein, 2018). These prototype models however solely focus on *between*-speaker variability, with each identity being conceptualized as a single point in space, neglecting to account for the substantial within-person variability. The findings reviewed above have shown that within-person variability is a key feature of human voices and there is some evidence that it affects voice identity perception. We argue therefore that it is important to empirically study the effects of within-person variability.

Effects of familiarity with a speaker on voice perception

Some authors have proposed that familiar and unfamiliar voice processing differs fundamentally from each other. For example, in their model of voice identity processing, Kreiman and Sidtis (2011) propose that unfamiliar voice perception relies on the comparison and discrimination of (acoustic) features in a voice (see also Van Lancker & Kreiman, 1987). In contrast to this, familiar voice perception is thought to rely on a more abstracted processing of identity, which can be achieved without explicit discrimination of a voice's acoustic features. Surprisingly, however, only a few studies have directly contrasted differences in identity processing for familiar and unfamiliar voices within the same task and, to date, a strong association between task type and listener characteristics is present in the existing literature. Studies have either employed voice recognition/identification tasks in the context of familiar voices (for an overview, see Kreiman & Sidtis, 2011) or used voice discrimination tasks with unfamiliar voices (Reich & Duke, 1979; Wester, 2012; Zarate *et al.*, 2015).

When directly comparing listener groups who are either familiar or unfamiliar with a set of test voices on a speaker discrimination task, a clear advantage for familiar listeners emerges (Lavan *et al.*, 2016). Complementary findings have also been reported for speech comprehension: Listeners are consistently better at understanding the speech of familiar voices compared to unfamiliar voices (Johnsrude, Casey, & Carlyon, 2014; Johnsrude *et al.*, 2013). Taken together, we can see general processing differences for familiar and unfamiliar voices, with advantages being apparent for extracting information from familiar voices.

Interactions of familiarity and within-person variability: insights from face perception

An issue that has not been extensively explored in the voice perception literature to date is the interaction of familiarity and within-person variability (but see Lavan *et al.*, 2016). Relevant insights on this topic may, however, be gleaned from the face perception literature, given the many proposed similarities between processing from faces and voices (Campanella & Belin, 2007; Kuhn, Wydell, Lavan, McGettigan, & Garrido, 2017; Yovel & Belin, 2013): Both signals convey a wealth of important information about a person, such as their age, sex, identity, emotions, and intentions. Furthermore, many parallels have been drawn between of how these kinds of information are processed in faces and voices – so much so, that the human voice has indeed been described as 'an auditory face' (Belin, Fecteau, & Bedard, 2004).

For face identity perception, stark differences in the processing of within-person variability for unfamiliar faces compared to familiar faces have been reported. We are able to reliably recognize familiar individuals even under challenging viewing conditions, for example, when images are degraded or include substantial within-person variability (Bruce, 1982; Hole, George, Eaves, & Rasek, 2002; Jenkins *et al.*, 2011; Yip & Sinha, 2002). With decreasing familiarity, our ability to tolerate such within-person variability also decreases (Bruce, Henderson, Newman, & Burton, 2001; Burton, Wilson, Cowan, & Bruce, 1999). In the absence of an abstracted representation of a face, variability across images, such as changes in viewpoint, expression, or lighting, or the type of camera used results in poor face identity matching and recognition for unfamiliar faces (Bruce, 1982; Bruce *et al.*, 2001; Bruce & Young, 1986; Henderson, Bruce, & Burton, 2001; Hill & Bruce, 1996; Kemp, Towell, & Pike, 1997).

These differences in how we cope with within-person variability in familiar and unfamiliar faces have been attributed to the nature of different representations available

for familiar and unfamiliar people (Bruce, 1982; Bruce & Young, 1986; Burton *et al.*, 2016; Hancock, Bruce, & Burton, 2000). While viewers have built up a relatively stable representation for a familiar face that is robust to changes in image properties, no such person-specific representations exist for unfamiliar faces. For unfamiliar faces, viewers are therefore thought to rely more on the visual properties of the specific unfamiliar face. These visual properties vary from image to image, resulting in the less accurate and more image-dependent perception of identity from unfamiliar faces.

A striking demonstration of the differences in the processing of identity in familiar and unfamiliar participants was provided by Jenkins *et al.* (2011) who used a face identity sorting task. Two groups of participants – one from the UK, the other from the Netherlands – sorted 40 images of two Dutch celebrities (20 images per identity) into piles by perceived identity. Crucially, these pictures were selected from Internet searches and thus included considerable within-person variability (different viewpoints, image quality, lighting, expressions, hairstyles, etc.). While participants from the Netherlands, who were familiar with these individuals, sorted the images most frequently into two piles (median = 2), participants from the UK, who were unfamiliar with the individuals, sorted the images most frequently into nine piles (median = 7.5). Despite perceiving more identities than the two that were actually present, unfamiliar participants only rarely sorted pictures of two different identities into the same pile. Unfamiliar participants were therefore able to successfully 'tell people apart', while they struggled to 'tell people together' and perceived the highly variable images from a single identity as several different identities.

This finding has since been replicated and extended: For example, the marked differences between familiar and unfamiliar viewers' behaviour have been shown to disappear when participants know how many identities to expect (Andrews, Jenkins, Cursiter, & Burton, 2015). Here, both viewer groups sorted the pictures into two piles with high accuracy and with few identity confusions. Redfern and Benton (2017) furthermore manipulated the expressiveness of unfamiliar faces, contrasting highly expressive versus less expressive (closer to neutral) faces in a sorting task. When faces were highly expressive, participants were more likely to sort two pictures from different identities into the same pile. Zhou and Mondloch (2016) showed an other-race effect in a face sorting task, where viewers sorted unfamiliar other-race faces into more perceived identities than unfamiliar own-race faces. This effect, however, was not present for familiar faces, where participants were highly accurate in both conditions. These face sorting studies thus show compelling interactions between familiarity and within-person variability where familiar individuals appear to be able to generalize across the variability, while unfamiliar individuals fail to do so in many cases.

The current study

Within-person variability is a key feature of human voices that listeners encounter in all of their everyday interactions with others. It has, however, to date been largely neglected in the study of voice perception – despite there being evidence that it affects voice identity perception. The face perception literature has shown that sorting tasks are a powerful tool for investigating different aspects of identity processing in the context of within-person variability (Jenkins *et al.*, 2011): Participants performance for 'telling people apart' and 'telling people together' can be assessed within a single task, while also being able to contrast performance for familiar versus unfamiliar voices.

In the current study, we investigated how within-person variability affects voice identity perception for familiar and unfamiliar voices using a voice sorting task. We selected voices from a popular TV show 'Orange is the New Black' and asked two groups of listeners, those who had watched the show (familiar listeners) and those who had not watched the show (unfamiliar listeners), to sort 30 voice samples (two voices, 15 exemplars per voice) into perceived identities. Crucially, our voice samples included natural within-person variability, having been extracted from different speaking situations, environments, and in the presence of different conversation partners (see Methods). We tested this voice sorting task in three independent participant samples, each using different stimulus sets to assess the replicability of effects. We predicted that unfamiliar listeners would perceive more identities than familiar listeners: In the absence of stable mental representation of a voice identity, natural within-person variability can be mistaken for between-person variability and can thus have a detrimental effect on accuracy. In terms of the composition of the formed clusters, we additionally predicted that unfamiliar listeners would be biased to mistaking within-person variability as between-person variability, thus selectively failing to 'tell people together' while being mostly able to 'tell people apart' (see Andrews *et al.*, 2015; Jenkins *et al.*, 2011; Redfern & Benton, 2017; Zhou & Mondloch, 2016 for faces).

Methods

Participants

A total of 152 participants were recruited via social media (e.g., Twitter and Facebook) and the participant pool of the Division of Psychology at Brunel University. Participants were either entered into a prize draw or received course credit for their participation. The study was approved by the local ethics committee. The 152 participants were randomly allocated to the three versions of the task (Sets 1–3; see below). Matching the sample size used by Jenkins *et al.* (2011), we aimed to recruit at least 20 participants for both our familiar and unfamiliar listener groups per set. Familiarity was assessed via self-report: If participants reported to have watched more than one season of 'Orange Is the New Black', they were assigned to the familiar group. Participants who reported to have not seen any episodes of the TV show were assigned to the unfamiliar group.[1] Participants who reported to have seen some episodes but not a full season were excluded from all analyses ($N = 7$). Participants, who reported that they had recognized or remembered more than three of the specific exemplars included in their set, were also excluded ($N = 3$) as their responses may have been driven by the specific memory of the scene as opposed to direct voice identity recognition. Additionally, we excluded participants who moved <80% of the exemplars (i.e., 24 exemplars out of 30; see below for information on the task) from their original position on the slide ($N = 1$) or whose performance (indexed by number of perceived identities; see below) differed by more than 3 standard deviations from the mean of their listener group and set ($N = 3$).

This resulted in a final data set of 68 familiar and 70 unfamiliar participants in total: 25 familiar (21 female, mean age: 18.68 years, *SD*: 1.15 years) and 22 unfamiliar participants (19 female, mean age: 18.70 years, *SD*: 1.72 years) for Set 1, 22 familiar (15 female,

[1] *This group assignment does not preclude the possibility that a subset of listeners labelled as 'unfamiliar' were nonetheless familiar with the voices by having watched other TV shows or films featuring these actors. We, however, note that the actors are currently primarily known for their performances in 'Orange is the New Black'.*

1 other, mean age: 24.36 years, *SD*: 7.78 years) and 22 unfamiliar participants (16 female, 1 other, mean age: 28.91 years, *SD*: 10.90 years) for Set 2, and 21 familiar (18 female, mean age: 26.48 years, *SD*: 4.12 years) and 26 unfamiliar participants (22 female, 1 other, mean age: 26.28 years, *SD*: 10.32 years) for Set 3.

Materials

We used exemplars of voices of three female characters with significant speaking roles from the TV show 'Orange Is the New Black' (VoiceID 1: Nicky Nichols, VoiceID 2: Alex Vause and VoiceID 3: Piper Chapman). The show was selected as it features many number of characters with significant speaking roles, providing a large pool of possible voices that could in principle be presented in the experiment.

Fifteen exemplars[2] per identity were extracted (mean duration: 3.12 s; *SD*: 0.32 s): These exemplars included full utterances with as little background noise as possible, avoiding catch phrases and other diagnostic verbal information (example stimulus: 'and that she is on her way out of town'). The linguistic content of the utterances differed from exemplar to exemplars and from identity to identity. To include substantial within-person variability in these samples, we ensured that each exemplar was extracted from a different scene, while the content of the utterance, speaking style, emotional content, and speaking environment was not controlled for and thus varied naturally (similar to the 'ambient images' in Jenkins *et al.*, 2011; see also Figure S1 for plots of affective and acoustic properties of the stimuli). Only recordings from the first three seasons of the TV show were included (released between 2 and 4 years before testing started) to decrease the likelihood that participants had recently heard the stimuli and would therefore remember the scenes in which they occurred. Exemplars were normed for intensity using PRAAT (Boersma & Weenink, 2017).

Procedure

There were three versions of the task (referred to as sets throughout the paper), including all possible pairs of the three different voices (Set 1: Nicky Nichols and Alex Vause, Set 2: Piper Chapman and Alex Vause, Set 3: Piper Chapman and Nicky Nichols) to assess the replicability of effects. Participants completed the experiment using the online testing platform Qualtrics (Qualtrics, Provo, UT, USA), where they downloaded a Microsoft PowerPoint slide that included 30 embedded sound files (2 identities $\times$ 15 exemplars). Each of these exemplars was represented by a number (see the bottom panel of Figure 1 for examples of listeners' completed solutions). Number and exemplar combinations were consistent across participants within set. The numbers were distributed evenly across the slide, with no clusters being obvious from the outset. In line with the methods used in Jenkins *et al.* (2011), participants were asked to sort the 30 exemplars into clusters, with each cluster including the exemplars produced by a single speaker, thus representing a perceived speaker identity. This was done via dragging and dropping the exemplars into clusters on the slide. Participants could replay the exemplars as many times as they wanted, and there was no time limit on completing the task.

[2] *Only 15 exemplars per identity were employed compared to the 20 exemplars used in Jenkins* et al.*'s (2011) study. We reasoned that there would be higher working memory demands when using voice compared to face stimuli, since participants cannot as readily compare exemplars in parallel and therefore reduced the number of exemplars.*

Results

Data were analysed both in terms of the number of perceived identities and how the exemplars were grouped, contrasting 'telling people apart' versus 'telling people together'. For all analyses, we reported the effects for each set separately, since the stimuli were different for each set. This then allowed us to assess consistency of effects across different stimuli. We used non-parametric tests throughout as Shapiro–Wilk tests showed that the data were not normally distributed in most conditions (i.e., for each set and listener group). α was Bonferroni-corrected for three comparisons for all analyses.

How many identities did familiar and unfamiliar listeners perceive?

For this analysis, we counted the number of clusters (i.e., how many identities listeners perceived) per participant. In two data sets (one familiar listener, one unfamiliar listener, both from Set 2), it was not clear whether one of the piles of exemplars formed by the participants was intended to represent one cluster or two clusters. In both cases, we counted these piles as two clusters. Familiar listeners perceived fewer clusters than unfamiliar listeners for all sets (see Figure 1, top panel. Familiar: Set 1 Median = 3, Mode = 2, Range = 2–8, Set 2 Median = 3, Mode = 2, Range = 2–8, Set 3 Median = 3, Mode = 2, Range = 2–12; Unfamiliar: Set 1 Median = 4, Mode = 4, Range = 2–10, Set 2 Median = 7, Mode = 5, Range = 4–17, Set 3 Median = 9, Mode = 11, Range = 3–15). Wilcoxon rank-sum tests confirmed that familiar listeners perceived significantly fewer identities than unfamiliar listeners for two out of the three individual sets (Set 1: $Z = 1.89$, $p = .030$; Set 2: $Z = 4.45$, $p < .001$; Set 3: $Z = 4.21$, $p < .001$).

In addition to differences in the number of perceived identities, we observed patterns of responses that were qualitatively different for familiar and unfamiliar listeners: Familiar listeners tended to create at least one – often two – large clusters (14+ items per cluster) plus a small number of single-exemplar clusters, resulting in a bimodal distribution of cluster sizes. Unfamiliar listeners, however, tended to create a number of smaller clusters (2–6 items, see Figure 1, middle panel). After collapsing all raw cluster counts across the three sets, a chi-squared test of independence confirmed that the distributions of frequencies of cluster sizes for familiar and unfamiliar listeners are independent, χ^2 (14) = 188.43, $p < .001$.

'Telling people apart' versus 'telling people together'

To assess the differences of familiar and unfamiliar listeners' ability to 'tell people apart' and conversely 'tell people together', we created 30 × 30 response matrices for each participant (15 sounds files × 2 identities; each cell shows the probability that two exemplars were sorted into the same cluster: Cells coded as 1 indicate that the two respective exemplars were always grouped together; cells coded as 0 indicate that the two exemplars were never grouped together — see Figure 2 bottom panel). These per participant response matrices thus provide a detailed representation of *how* listeners grouped the different sounds into perceived identities. We used these matrices to characterize errors in 'telling people apart' and 'telling people together'. Figure 2 (top panel) shows the group-averaged response matrices. Conceptually, these matrices are divided into within-identity and across-identity submatrices (see Figure 2, bottom panel). Within-identity submatrices index listeners' ability to 'tell people together': For the ideal solution (creating the two correct clusters), each cell within these submatrices would be 1 as all pairs of exemplars from the same identity were put into the same cluster. The

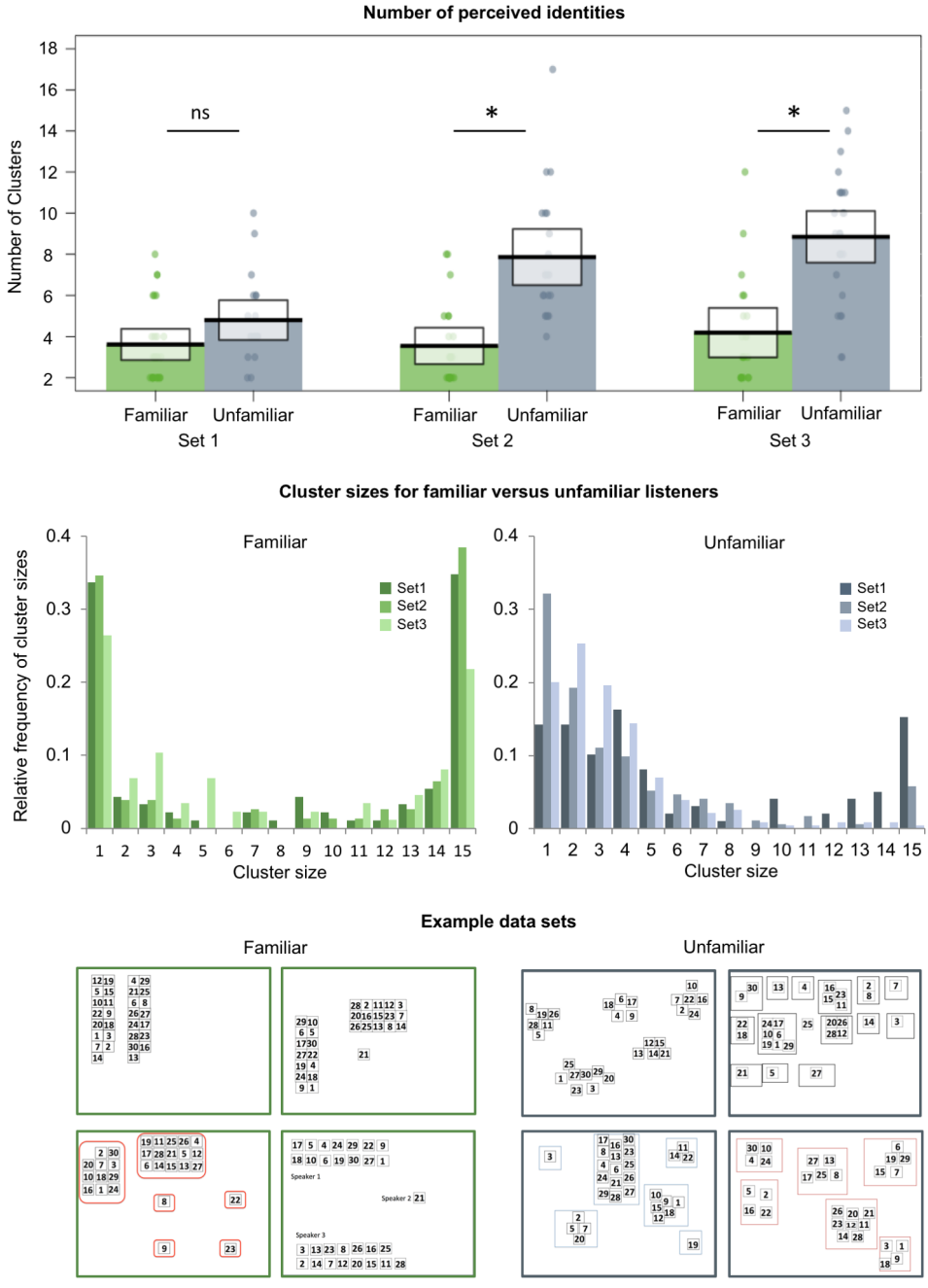
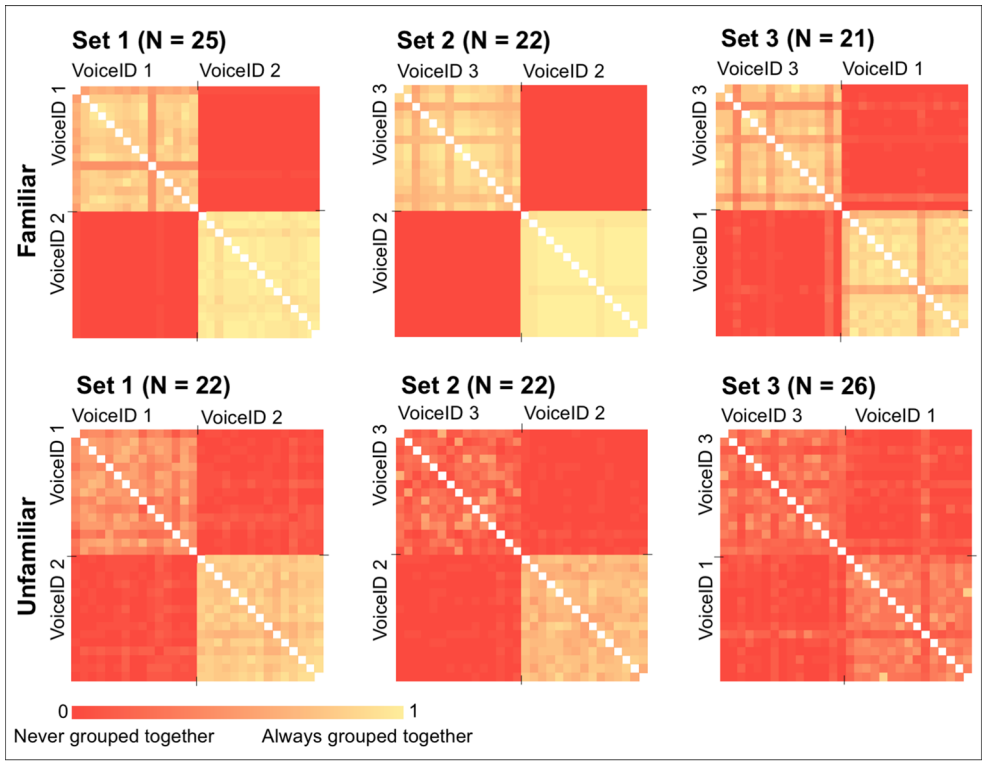


Figure 1. Top panel: Number of perceived identities by set for familiar and unfamiliar listeners. Bars show the means across participants, and each dot shows one participant's data. Boxes show the 95% confidence intervals for the means of each. Stars show significant differences between familiar and unfamiliar listeners (α was Bonferroni-corrected for three comparisons). Middle panel: Plots of the relative frequency of cluster sizes (count per cluster size divided by the total number of clusters within each set) for familiar and unfamiliar listeners. Bottom panel: Representative example data sets as an illustration of familiar and unfamiliar participants' response patterns. [Colour figure can be viewed at wileyonlinelibrary.com]



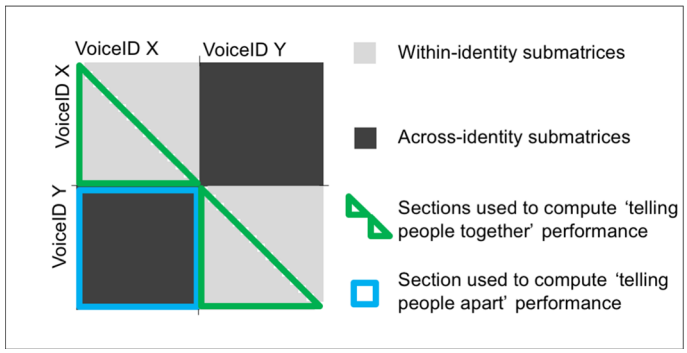


Figure 2. Top panel: Matrices of averaged listeners' responses for the three versions of the task for familiar and unfamiliar listeners. Within these 30 × 30 matrices (15 sounds files × 2 identities), each cell shows the probability with which two exemplars were grouped within the same perceived identity: Cells with a value of 1 indicate that the respective exemplars were always clustered together, cells with a value of 0 indicate that these sounds were never in the same clusters. Bottom panel: Illustration of the different sections of the per participant matrices that were analysed below. [Colour figure can be viewed at wileyonlinelibrary.com]

across-identity submatrix indexes the ability to 'tell people apart': An ideal solution here would result in all cells within this submatrix to be 0 as no pairs of exemplars from different identities were ever put into the same cluster (see Figure 2, bottom panel, the marked submatrices only cover one half of the matrices as they are by definition symmetrical across the diagonal).

To quantify whether familiar and unfamiliar listeners' performance for 'telling people together' and 'telling people apart' differed, we computed the mean probability of how often two exemplars from the same identity were grouped together by taking the mean of the values in the lower triangle of each symmetrical within-identity submatrix (excluding the diagonal which is by definition always 1 and therefore not meaningful; 2 × 105 cells, see Figure 2 bottom panel). The median probabilities for 'telling people together' were higher for familiar listeners (Set 1 = .88, Set 2 = .93, Set 3 = .81) than for unfamiliar listeners (Set 1 = .64, Set 2 = .46, Set 3 = .18). Wilcoxon rank-sum tests confirmed that familiar listeners were significantly more likely to group exemplars from the same identity together than unfamiliar listeners for all three sets (all Zs > 3.41, all ps < .001).

A comparable analysis was run for 'telling exemplars apart' submatrices. We computed the mean of the values in the across-identity matrix (225 cells, see Figure 2 bottom panel). The median value for 'telling people apart' was overall very low (or 0) for familiar listeners (Set 1 = 0, Set 2 = 0, Set 3 = .02) as well as unfamiliar listeners (Set 1 = 0, Set 2 = 0, Set 3 = .04). Wilcoxon rank-sum test showed that unfamiliar listeners made significantly more errors than familiar listeners in only one of the three sets (Set1: $Z = 2.67$, $p = .004$; Set 2: $Z = 1.77$, $p = .038$; Set 3: $Z = 1.48$, $p = .069$).

Explicit comparisons of error rates (i.e., we computed 1 minus the mean probability of the lower triangle of each within-identity matrix for each participant and compared it with the mean probability of the across-identity matrices) showed that there were indeed significant differences for familiar and unfamiliar listeners for all sets (familiar: all Zs > 2.95, all ps < .002; unfamiliar: all Zs > 3.94, all ps < .001). 'Telling people together' can thus be considered a more challenging or error-prone process.

Overall, striking differences in the behaviour of familiar and unfamiliar listeners are apparent: Unfamiliar listeners were less likely to group exemplars from the same identity together compared to familiar listeners. In contrast, both listener groups largely succeeded at telling the two different identities apart, given the very low error rates.

How similar are individual response matrices to each other?

We have so far shown that response matrices differ from each other in a number of ways across familiar and unfamiliar listeners. We next explored whether the response patterns of individual listeners within a group differ from each other or are highly similar. Each participant's 30 × 30 response matrix was correlated with every other participant's matrices within their set and listener group using Kendall's τ_a. We obtained a mean correlation per participant and then computed the mean across participants.

These analyses showed that the matrices for all three sets and both listener groups were significantly correlated (Familiar Mean Kendall's τ_a: Set 1 = .359, Set 2 = .386, Set 3 = .269; Unfamiliar Mean Kendall's τ_a: Set 1 = .178, Set 2 = .148, Set 3 = .032; Wilcoxon's signed rank tests against 0, familiar: all Zs > 3.83, all ps < .001; unfamiliar: all Zs > 3.89, all ps < .001). Mean correlations were significantly stronger among familiar listeners compared to unfamiliar listeners for all sets (Wilcoxon's rank-sum tests, all Zs > 5.66, all ps < .001). These results show that familiar listeners arrived at more similar solutions compared to unfamiliar listeners, probably due to better task performance (i.e., the number of perceived identities was closer to the veridical number of identities present). While some consistency is present in the ratings of the unfamiliar listeners, participants seem to have arrived at quite dissimilar solutions (most strikingly illustrated in Set 3). This may indicate that there are a number of different strategies to complete the task

(see Data S1 for an analysis attempting to link response pattern to acoustic and perceptual properties of the exemplars).

Not all voices are alike: Effects of different speaker identities and context

The group-averaged matrices show that there were identity-specific effects, with exemplars for some identities being easier to 'tell together' than for other identities (see Figure 2, top panel). We tested whether the probabilities with which listeners grouped the exemplars of an identity together differed for the three voices, using within-set comparisons (e.g., comparing VoiceID1 from Set 1 vs. VoiceID2 from Set 1). Wilcoxon signed rank tests showed that the probability of 'telling people together' for the same identity differed across most voices for familiar listeners (VoiceID1 vs. VoiceID2; VoiceID2 vs. VoiceID3: both Zs > 2.99; ps < .002). The comparison of VoiceID1 versus VoiceID3 did not reach significance (Z = .40; p = .665). These results show that familiar listeners were more successful at telling the exemplars of VoiceID2 together compared to the other two identities, marking this particular voice as being potentially more distinctive or less inherently variable than the other voices (see also Figure 2). This is also reflected in the performance of unfamiliar listeners: performance here was highest for VoiceID2, with an additional difference emerging between performance for VoiceID1 and VoiceID3 (all Zs > 2.58; all ps < .005).

We also examined effects of context by testing whether the probability with which listeners grouped different exemplars of an identity together differed depending on the other identity included alongside (e.g., VoiceID1 from Set 1 vs. VoiceID1 from Set 3; see above for methods). Interestingly, the probabilities did not differ from each other for familiar listeners for any of the voices (Wilcoxon rank-sum tests; all Zs < .55, all ps > .399). For unfamiliar listeners, probabilities only differed for VoiceID1, showing better performance in Set 1 compared to Set 3 (VoiceID1: Z = 2.66, p = .004, VoiceID2 and VoiceID3: Zs < .81, ps > .210). This result shows that the nature of the other identity included in the sets (or indeed the participant sample) did in the main not significantly affect how difficult it was to tell exemplars of the same identity together (except for VoiceID1 for unfamiliar listeners) and thus speaks against consistent effects of context. It remains unclear why performance was significantly different between sets for VoiceID1 only for unfamiliar listeners. This finding however sheds some light on how Set 1 differs from the remaining two sets, being the only set that did not show a statistically significant difference between the number of clusters formed by familiar versus unfamiliar listeners.

Not all exemplars are alike: effects of perceived likeness

Not all voices are alike, but not all exemplars may be alike either. In the context of within-person variability, some exemplars can sound more like a familiar person than others (Ritchie, Kramer, & Burton, 2018, for faces). To investigate whether perceived likeness has an effect on how identity information is processed, we computed the mean cluster size for each exemplar, averaged across the two instances in which each exemplar occurred across the three sets. Our previous analyses have shown that familiar and unfamiliar listeners generally succeed at 'telling identities apart' but unfamiliar listeners struggle to 'tell identities together', resulting in a larger number of perceived identities. In this context, cluster size per exemplar can thus serve as an index of how difficult listeners found it to associate a particular exemplar with the other exemplars of this identity.

We collected perceptual ratings of perceived likeness for each exemplar from an independent group of 15 listeners who were familiar with the TV show (13 female; mean age = 19.06 years, SD = 0.77 years) at Royal Holloway, University of London. They

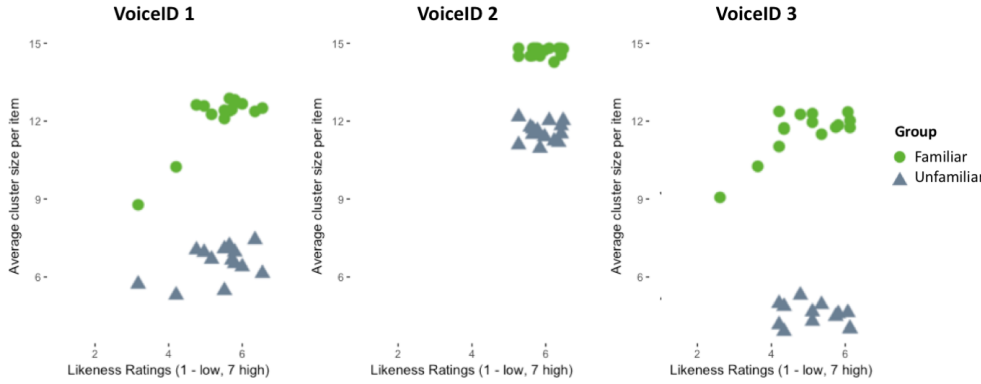


Figure 3. Scatter plots of the exemplar-wise mean cluster size and likeness ratings per identity for familiar and unfamiliar listeners. Cluster size was averaged across the two samples in different sets. [Colour figure can be viewed at wileyonlinelibrary.com]

received course credit for their participation. The study was approved by the local ethics committee. Participants were presented with the 45 exemplars (15 × 3 identities), blocked by speaker identity. The order of identity blocks and order of stimuli within each block were randomized. Participants provided ratings of perceived likeness on a scale from 1 to 7 ('How much does this sound like [character name]?'; 1 = not at all, 7 = very much). They were asked to rate the quality of the voice and disregard the verbal content of the stimuli. From these ratings, mean ratings of likeness were computed per exemplar. Likeness ratings for VoiceID1 were lost from one participant due to a technical error. No participants were excluded. Figure 3 illustrates the substantial variability in mean perceived likeness for different items of the same speaker.

To explore the relationship between average cluster size (based on the data from the main sorting tasks) and the mean perceived likeness ratings collected from an independent group of familiar listeners, we computed Kendall's τ_a between the two measures. This was done separately for average cluster size measures for unfamiliar and familiar listeners and for each voice. Significance was determined through random permutation tests (5,000 iterations). If the observed value of Kendall's τ_a was higher than 95% of the chance predictions ($p < .05$) obtained by shuffling the values within the comparisons of interest, we rejected the null hypothesis.

For familiar listeners, correlations were not significant after correcting for multiple comparisons for VoiceID1 (Kendall's $\tau_a = .35$, $p = .035$) and VoiceID3 (Kendall's $\tau_a = .32, p = .040$). The correlation for VoiceID2 was not significant (Kendall's $\tau_a = .095$, $p = .298$) due to a ceiling effect for this particular voice (see Figure 3). The correlations for unfamiliar listeners were not significant for any of the voice identities (Kendall's $\tau_a < .120, p > .277$). No definitive relationship between perceived likeness and identity processing can thus be established from the current data, although trends are apparent for familiar listeners. Here, the items with lowest ratings of likeness were, however, also clearly the items with the smallest average cluster size (see Figure 3).

Discussion

The current study explored how *natural* within-person variability affects voice identity processing in familiar and unfamiliar listeners within the same paradigm. When asked to

group 30 sound clips from a popular TV show (two voices, 15 exemplars each) into perceived identities, familiar listeners perceived on average between three and four speakers. In contrast, unfamiliar listeners perceived more speakers (on average between four and nine speakers). Thus, unfamiliar listeners perceived numerically more identities across all three sets compared to familiar listeners, although this difference was only significant for Sets 2 and 3. This discrepancy can be partially explained by a context (or sample) effect apparent for VoiceID1 in Set 1: For this specific voice, unfamiliar listeners performed significantly better in Set 1 compared to the listeners in Set 3, lowering the overall number of clusters in this particular set. Across sets and listener groups, substantial individual differences are furthermore apparent: For familiar listeners, this may reflect the inclusion of participants with a varied duration of exposure (between one and five seasons with variable amounts of time having passed since watching the show) and engagement with the show. Furthermore, these differences could also reflect more general individual differences in voice identity processing (see Aglieri *et al.*, 2017).

In terms of how the clusters were formed, our results show that unfamiliar listeners frequently perceived exemplars from the same speaker as different identities pointing to selective difficulties in 'telling people together' by failing to successfully generalize identity information across variable signals. Both listener groups only made a relatively small number of errors in 'telling people apart' by grouping exemplars from two identities into the same cluster. These findings are thus a first direct demonstration of unfamiliar listeners' failure to 'tell people together' in the context of naturally varying voice recordings (for comparable findings for faces, see Jenkins *et al.*, 2011) and highlight the need to consider within-person variability, a feature central to human voices, in models and studies of voice identity perception.

We further explored whether there were effects of specific voices and items. Identities did indeed differ in their overall difficulty for both familiar and unfamiliar listeners groups. These differences between voices may reflect the fact that some voices may be inherently more distinctive than others to most listeners or that they may be inherently less variable. Context, provided by the second voice within a pair, on the other hand had no consistent effect on listeners' judgements of telling exemplars of an identity together in the current study (but see an effect of context for VoiceID1 for unfamiliar listeners). Additionally, we observed exemplar effects: Here, not all exemplars were equally easy to group with the other exemplars produced by the same speaker. In other words, not all exemplars were equally easy to 'tell together' (as indexed by the mean cluster size per exemplar). An analysis investigating the link between perceived likeness and listeners' ability to group items together did not reveal any statistically significant results. For familiar listeners, interesting trends emerged, indicating that exemplars rated to be a relatively 'bad likeness' of a person may be difficult to associate with other exemplars. These trends may indicate that familiar listeners' performance can be systematically affected by certain aspects of within-person variability, indexed here by differences in perceived likeness. Further studies are, however, required to fully explore this potential relationship.

While current models of voice processing do not explicitly account for within-person variability and only little empirical evidence probing this issue is available to date, the findings of our study can nonetheless be integrated into and advance current models of voice processing. The model of voice identity processing proposed by Kreiman and Sidtis (2011; Sidtis & Kreiman, 2012) focuses on the distinction of familiar and unfamiliar voice processing during identity perception. Here, familiar voice recognition and unfamiliar

voice discrimination are considered to be mechanistically distinct (featural comparison vs. pattern recognition), are dissociable from one another, and thus predict differences in the behaviour of familiar and unfamiliar listeners. In our study, we indeed found striking differences between familiar and unfamiliar listeners' performance, using the same task for both listener groups (note, however, that we failed to link performance for unfamiliar listeners to discrete acoustic features, see Data S1).

Prototype models of voice processing may offer some insights into the nature of the different representations of familiar and unfamiliar voices. Such prototype models propose that listeners encode and process voice identity information in relation to a prototype, which is a context-dependent average voice (Latinus & Belin, 2011; Latinus *et al.*, 2013; Lavner *et al.*, 2001; Papcun *et al.*, 1989; see also Maguinness *et al.*, 2018). While empirical studies show some support for these models, these studies have to our knowledge only explored prototype models with a focus on *between*-speaker variability by using different voice identities (see Lavan *et al.*, 2018, for a discussion). The mechanisms assumed for prototype models can, nonetheless, be readily extended and applied to the processing of *within*-person variability: For a familiar voice, listeners can access a specific prototype or representation of a particular voice. These representations of familiar voices are likely to include the characteristics of how a specific voice varies. Due to this, listeners can thus still relatively reliably process voice identity from known voices, even in the face of within-person variability. For unfamiliar voices, neither a specific representation is available nor have the characteristics of how a specific voice varies been encoded. The lack of specific information may thus result in the processing of identity being less reliable.

Our study's findings closely resemble the results reported for faces (Jenkins *et al.*, 2011). Many parallels have in the past been described between face and voice processing (Campanella & Belin, 2007; Kuhn *et al.*, 2017; Yovel & Belin, 2013). The degree of similarity of results between the current auditory sorting task and visual identity sorting studies is nonetheless remarkable, given the differences in the materials used in face and voice sorting studies. Not only do the materials derive from two different modalities, they also provide participants with in the case of faces with static information and while they provide dynamic information in the case of voices. Given these profound differences in the signals, the nature of the within-person variability present in both sets will also differ accordingly. There is likely no clear one-to-one correspondence between the sources of variability: How does, for example, variability in the lighting of images relate to variability introduced by background noise? There is also no direct equivalent for differences in viewpoint in the auditory domain nor can we adequately describe a regional accent in the (static) visual domain. In short, salient features for identity processing and sources of variability are likely to be modality-specific.

Aside from differences in materials, the task of identity sorting allows participants to choose their own strategy to complete the tasks with no explicit instructions guiding them. These strategies may differ between faces and voice versions of the task based on the nature of the stimuli. For example, a voice sorting task is more demanding on working memory: During a face sorting task, the image never disappears, while the voice disappears as soon as the playback stops. Listeners thus need to at least partially memorize items. Despite these factors, patterns of results for face and voice sorting tasks are comparable: Such parallels may suggest that sorting tasks tap into stages of identity processing in familiar as well as unfamiliar participants that may either rely on abstracted amodal processes or alternatively modality-bound

processes that are mirrored closely in the auditory and visual domain (see Yovel & Belin, 2013). Mapping out in which contexts the processing of face and voice identities is comparable and in which circumstances the two modalities differ remains a largely open question and warrants further work.

One of the novel aspects of this study is its use of the relatively uncontrolled exemplars that include substantial, natural within-person variability (similar to the 'ambient images' of faces used in Jenkins *et al.*, 2011): Exemplars varied in extrinsic features, such as the overall quality of the recording, type and amount of background noise among any number of other factors. Furthermore, exemplars differed in their linguistic/verbal content (different utterances within and across voice identity), verbal register, type of utterance, vocal effort (quiet conversation vs. shouting) as well as their perceived affective properties, such as valence and arousal, and perceived likeness, among any number of features. While the current study shows how uncontrolled natural within-person variability from a range of sources can affect speaker identity perception, other studies have shown how specific sources of variability can affect perception (e.g., language spoken, Zarate *et al.*, 2015; linguistic content, Narayan *et al.*, 2017; vocalizations type, Lavan *et al.*, 2016; distinctiveness, Papcun *et al.*, 1989; and duration of the exemplars Schweinberger, Herholz, & Sommer, 1997). How these different types of variability relate to each other and interact in the context of identity perception is largely unexplored. Similarly, we do not know whether different types of variability (e.g., variability introduced by voice modulations vs. variability introduced by recording quality) might be more disruptive to perception than others or whether their effects are comparable to each other. Further studies are therefore needed to better characterize the nature of within-person variability and its effects on identity perception.

The present study thus demonstrates that within-person variability poses challenges for the reliable processing of identity from voices – especially for unfamiliar listeners. Within-person variability may, however, not always be a challenge that listeners need to overcome as recent intriguing findings from the face identity perception literature suggest. Burton *et al.* (2016) showed that within-person variability is specific to an individual's face, that is *how* the face of one person varies is different from how another face varies. Within-person variability may therefore encode diagnostic information about a person's identity, as opposed to merely being noise. There is also some evidence that within-person variability may indeed be instrumental to building up robust representations of a person, given that participants are more successful at learning a novel identity from training with variable sets of face stimuli compared to when trained on less variable sets (Murphy, Ipser, Gaigg, & Cook, 2015; Ritchie & Burton, 2017). Given the striking parallels between the findings of the current study and reports from face sorting tasks, it is possible that the processing proposed for identity learning from variable faces may also extend to how voices are learnt. Future work will therefore not only need to map out *how* listeners' judgements are affected by within-person variability, but will also need to explore whether and how within-person variability could be an essential part of voice identity learning.

Acknowledgements

This work was supported by a research project grant from the Leverhulme Trust (RPG-2014-392) awarded to Lúcia Garrido. We would like to thank Ibtisam Abdi and Saira Mahmood Khan for help with the data entry and Matthew Longo for comments on a draft of this manuscript.

References

Aglieri, V., Watson, R., Pernet, C., Latinus, M., Garrido, L., & Belin, P. (2017). The Glasgow Voice Memory Test: Assessing the ability to memorize and recognize unfamiliar voices. *Behavior Research Methods*, *49*(1), 97–110. https://doi.org/10.3758/s13428-015-0689-6

Andrews, S., Jenkins, R., Cursiter, H., & Burton, A. M. (2015). Telling faces together: Learning new faces through exposure to multiple instances. *The Quarterly Journal of Experimental Psychology*, *68*, 2041–2050. https://doi.org/10.1080/17470218.2014.1003949

Belin, P., Fecteau, S., & Bedard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences*, *8*(3), 129–135. https://doi.org/10.1016/j.tics.2004.01.008

Boersma, P., & Weenink, D. (2017). *Praat: Doing phonetics by computer [Computer program]*. Retrieved from http://www.praat.org/

Bruce, V. (1982). Changing faces: Visual and non-visual coding processes in face recognition. *British Journal of Psychology*, *73*(1), 105–116. https://doi.org/10.1111/j.2044-8295.1982.tb01795.x

Bruce, V., Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied*, *7*(3), 207. https://doi.org/10.1037/1076-898X.7.3.207

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, *77*, 305–327. https://doi.org/10.1111/j.2044-8295.1982.tb01795.x

Burton, A. M. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *The Quarterly Journal of Experimental Psychology*, *66*, 1467–1485. https://doi.org/10.1080/17470218.2013.800125

Burton, A. M., Kramer, R. S., Ritchie, K. L., & Jenkins, R. (2016). Identity from variation: Representations of faces derived from multiple instances. *Cognitive Science*, *40*(1), 202–223. https://doi.org/10.1111/cogs.12231

Burton, A. M., Wilson, S., Cowan, M., & Bruce, V. (1999). Face recognition in poor-quality video: Evidence from security surveillance. *Psychological Science*, *10*, 243–248. https://doi.org/10.1111/1467-9280.00144

Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences*, *11*, 535–543. https://doi.org/10.1016/j.tics.2007.10.001

Hancock, P. J., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences*, *4*, 330–337. https://doi.org/10.1016/S1364-6613(00)01519-9

Henderson, Z., Bruce, V., & Burton, A. M. (2001). Matching the faces of robbers captured on video. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *15*(4), 445–464.

Hill, H., & Bruce, V. (1996). Effects of lighting on the perception of facial surfaces. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 986–1004. https://doi.org/10.1037/0096-1523.22.4.986

Hole, G. J., George, P. A., Eaves, K., & Rasek, A. (2002). Effects of geometric distortions on face-recognition performance. *Perception*, *31*, 1221–1240. https://doi.org/10.1068/p3252

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition*, *121*, 313–323. https://doi.org/10.1016/j.cognition.2011.08.001

Johnsrude, I., Casey, E., & Carlyon, R. P. (2014). Listen to your mother: Highly familiar voices facilitate perceptual segregation. *The Journal of the Acoustical Society of America*, *135*, 2423. https://doi.org/10.1121/1.4878052

Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P., & Carlyon, R. P. (2013). Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice. *Psychological Science*, *24*, 1995–2004. https://doi.org/10.1177/0956797613482467

Kemp, R., Towell, N., & Pike, G. (1997). When seeing should not be believing: Photographs, credit cards and fraud. *Applied Cognitive Psychology*, *11*, 211–222. https://doi.org/10.1002/(SICI)1099-0720(199706)11:3<211:AID-ACP430>3.0.CO;2-O

Kreiman, J., Park, S. J., Keating, P. A., & Alwan, A. (2015). The relationship between acoustic and perceived intraspeaker variability in voice quality. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Kreiman, J., & Sidtis, D. (2011). *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. Chichester, UK: Wiley-Blackwell.

Kuhn, L. K., Wydell, T., Lavan, N., McGettigan, C., & Garrido, L. (2017). Similar representations of emotions across faces and voices. *Emotion*, *17*, 912–937. https://doi.org/10.1037/emo0000282

Latinus, M., & Belin, P. (2011). Anti-voice adaptation suggests prototype-based coding of voice identity. *Frontiers in Psychology*, *2*, 175. https://doi.org/10.3389/fpsyg.2011.00175

Latinus, M., McAleer, P., Bestelmeyer, P. E., & Belin, P. (2013). Norm-based coding of voice identity in human auditory cortex. *Current Biology*, *23*, 1075–1080. https://doi.org/10.1016/j.cub.2013.04.055

Lavan, N., Burton, A. M., Scott, S. K., & McGettigan, C. (2018). Flexible voices: Identity perception from variable vocal signals. *Psychonomic Bulletin & Review*, 1–13. https://doi.org/10.3758/s13423-018-1497-7

Lavan, N., Scott, S. K., & McGettigan, C. (2016). Impaired generalization of speaker identity in the perception of familiar and unfamiliar voices. *Journal of Experimental Psychology: General*, *145*, 1604–1614. https://doi.org/10.1037/xge0000223

Lavner, Y., Rosenhouse, J., & Gath, I. (2001). The prototype model in speaker identification by human listeners. *International Journal of Speech Technology*, *4*(1), 63–74. https://doi.org/10.1023/A:1009656816383

Maguinness, C., Roswandowitz, C., & Von Kriegstein, K. (2018). Understanding the mechanisms of familiar voice-identity recognition in the human brain. *Neuropsychologia*, *116*, 179–193. https://doi.org/10.1016/j.neuropsychologia.2018.03.039

Murphy, J., Ipser, A., Gaigg, S. B., & Cook, R. (2015). Exemplar variance supports robust learning of facial identity. *Journal of Experimental Psychology: Human Perception and Performance*, *41*, 577–581. https://doi.org/10.1037/xhp0000049

Narayan, C. R., Mak, L., & Bialystok, E. (2017). Words get in the way: Linguistic effects on talker discrimination. *Cognitive Science*, *41*, 1361–1376. https://doi.org/10.1111/cogs.12396

Papcun, G., Kreiman, J., & Davis, A. (1989). Long-term memory for unfamiliar voices. *The Journal of the Acoustical Society of America*, *85*, 913–925. https://doi.org/10.1121/1.397564

Peynircioğlu, Z. F., Rabinovitz, B. E., & Repice, J. (2017). Matching speaking to singing voices and the influence of content. *Journal of Voice*, *31*, 256.e13–256.e17. https://doi.org/10.1016/j.jvoice.2016.06.004

Read, D., & Craik, F. I. (1995). Earwitness identification: Some influences on voice recognition. *Journal of Experimental Psychology: Applied*, *1*(1), 6–18. https://doi.org/10.1037/1076-898X.1.1.6

Redfern, A. S., & Benton, C. P. (2017). Expressive faces confuse identity. *i-Perception*, *8*, https://doi.org/10.1177/2041669517731115

Reich, A. R., & Duke, J. E. (1979). Effects of selected vocal disguises upon speaker identification by listening. *The Journal of the Acoustical Society of America*, *66*(4), 1023–1028. https://doi.org/10.1121/1.383321

Ritchie, K. L., & Burton, A. M. (2017). Learning faces from variability. *The Quarterly Journal of Experimental Psychology*, *70*, 897–905. https://doi.org/10.1080/17470218.2015.1136656

Ritchie, K. L., Kramer, R. S., & Burton, A. M. (2018). What makes a face photo a 'good likeness'? *Cognition*, *170*, 1–8. https://doi.org/10.1016/j.cognition.2017.09.001

Saslove, H., & Yarmey, A. D. (1980). Long-term auditory memory: Speaker identification. *Journal of Applied Psychology*, *65*(1), 111–116. https://doi.org/10.1037/0021-9010.65.1.111

Schweinberger, S. R., Herholz, A., & Sommer, W. (1997). Recognizing famous voices: Influence of stimulus duration and different types of retrieval cues. *Journal of Speech, Language, and Hearing Research*, *40*, 453–463. https://doi.org/10.1044/jslhr.4002.453

Sidtis, D., & Kreiman, J. (2012). In the beginning was the familiar voice: Personally familiar voices in the evolutionary and contemporary biology of communication. *Integrative Psychological and Behavioral Science*, *46*, 146–159. https://doi.org/10.1007/s12124-011-9177-4

Van Lancker, D., & Kreiman, J. (1987). Voice discrimination and recognition are separate abilities. *Neuropsychologia*, *25*, 829–834. https://doi.org/10.1016/0028-3932(87)90120-5

Wagner, I., & Köster, O. (1999). Perceptual recognition of familiar voices using falsetto as a type of voice disguise. *Proceedings of the 14th International Congress of Phonetic Sciences*, *2*, 1381–1384. https://doi.org/10.1016/0028-3932(87)90120-5

Wester, M. (2012). Talker discrimination across languages. *Speech Communication*, *54*, 781–790. https://doi.org/10.1016/j.specom.2012.01.006

Yip, A. W., & Sinha, P. (2002). Contribution of color to face recognition. *Perception*, *31*, 995–1003. https://doi.org/10.1068/p3376

Yovel, G., & Belin, P. (2013). A unified coding strategy for processing faces and voices. *Trends in Cognitive Sciences*, *17*, 263–271. https://doi.org/10.1016/j.tics.2013.04.004

Zarate, J. M., Tian, X., Woods, K. J., & Poeppel, D. (2015). Multiple levels of linguistic and paralinguistic features contribute to voice recognition. *Scientific Reports*, *5*, 11475. https://doi.org/10.1038/srep11475

Zhou, X., & Mondloch, C. J. (2016). Recognizing "Bella Swan" and "Hermione Granger": No own-race advantage in recognizing photos of famous faces. *Perception*, *45*, 1426–1429. https://doi.org/10.1177/0301006616662046

Received 2 May 2018; revised version received 12 August 2018

Supporting Information

The following supporting information may be found in the online edition of the article:

Data S1. Supplementary analyses.
Figure S1. Matrices for candidate models for the three sets.
Figure S2. Plots of arousal and valence ratings and acoustic features.